https://doi.org/10.54141/psbd.1489554

PAMUKKALE JOURNAL OF SPORT SCIENCES
ISSN: 1309-0356
dergipark.org.tr/en/pub/psbd

Pamukkale J Sport Sci, 15(3), 455-485, 2024

Research Article

The Role of Performance Metrics in Estimating Market Values of Footballers in Europe's Top Five Leagues

Mehmet Ali YALÇINKAYA[1*] Murat IŞIK[1]

[1]*Department of Computer Engineering, Faculty of Engineering and Architecture, Kırşehir Ahi Evran University, Kırşehir, Türkiye*

Keywords
Football player valuation,
Performance metrics,
Transfer market analysis,
Machine learning in sports

Article History
Received 24 May 2024
Revised 30 September 2024
Accepted 30 October 2024
Available Online 30 December 2024

* **Corresponding Author**:
Mehmet Ali YALÇINKAYA
E-mail Address :
mehmetyalcinkaya@ahievran.edu.tr

ABSTRACT

The transfer economy in football is a multi-billion-dollar industry, where accurate valuation of players is crucial for clubs' financial sustainability and competitive success. This study investigates the role of performance metrics in estimating the market values of football players in Europe's top five leagues (Spain's La Liga, France's Ligue 1, England's Premier League, Italy's Serie A, and Germany's Bundesliga). The study collected 28 performance metrics (e.g., goals, shots per game, assists, and pass success percentage) for 1508 players from the Whoscored platform. Additionally, the players' positions and the leagues they play in were also included as features. These data were combined with market values from the Transfermarkt platform, resulting in a comprehensive dataset. Two main analytical methods were employed: regression and classification. In the regression analysis, seven models (Adaboost, Decision Tree, Gradient Boosting, K Nearest Neighbors, Random Forest, Ridge Regression, and Support Vector Machine) predicted players' market values. The highest accuracy was achieved with the Random Forest algorithm (R-squared: 0.90). In the classification analysis, players' market values were categorized into four classes (low, lower-mid, upper-mid, and high), and their class memberships were predicted based on performance metrics. The CNN algorithm achieved the highest accuracy, with a success rate of 97%. The results indicate that performance metrics significantly contribute to estimating football players' market values, and models based on these metrics can assist clubs in making more informed, data-driven decisions during transfers.

Citation: Yalçınkaya, M. A. and Işık, M. (2024). The Role of Performance Metrics in Estimating Market Values of Footballers in Europe's Top Five Leagues. *Pamukkale Journal of Sport Sciences, 15*(3), 455-485. *https://doi.org/10.54141/psbd.1489554*

INTRODUCTION

Football, a sport with roots in ancient civilizations, attained its modern form in 19th-century England. The establishment of the Football Association in 1863 marked the beginning of standardized rules, which soon spread globally (Murray, 1994). The rapid globalization of football transformed it into the world's most popular sport in a short period. Today, football plays a significant role in bringing together millions of people from diverse cultures and fostering social and cultural bonds (Goldblatt, 2006).

Beyond being a mere sport, football is a force that shapes social dynamics and unites communities. Major tournaments enhance feelings of friendship and competition among nations. Football's impact on the masses is evident in the deserted streets on match days, the packed stadiums, and the millions of viewers in front of screens. Consequently, football is not just a sporting event but a social phenomenon (Giulianotti, 1999).

The economic dimension of football has undergone dramatic changes over time. While football was considered an amateur sport at the beginning of the 20th century, it has since evolved into a multi-billion-dollar industry. The formation of professional leagues and the sale of television broadcast rights significantly increased the economic value of football. Since the 1990s, football clubs have gained substantial financial power through sponsorship deals, ticket sales, merchandise, and media rights.

Enormous expenditures made by clubs mark the football transfer market to enhance their competitive edge. According to the Deloitte Football Money League (2024) report, the total annual revenue of European football reached €10.5 billion for the 2023/24 season (Deloitte, 2024). This figure illustrates that football is not just a sport but a colossal economic sector. This growth has enhanced football's integration with the media, advertising, and entertainment industries, boosting its economic potential.

In recent years, the astronomical amounts spent on football transfers have become evident, exemplified by Neymar's 2017 transfer to Paris Saint-Germain for €222 million (Bida & Mirzoyan, 2023). Such high-profile transfers attract the attention of clubs' fans and the media. However, these expenditures also raise concerns about clubs' financial sustainability.

The substantial transfer expenditures and financial management issues faced by football clubs have led some to bankruptcy. For instance, in 2012, Glasgow Rangers went into administration due to financial troubles and were demoted to lower leagues (BBC Sport, 2012). Similarly, in 2015, Parma FC in Italy went bankrupt due to a financial crisis and was relegated to Serie D (Gazzetta dello Sport, 2015).

Clubs' financial challenges are not solely due to poor financial management. Many clubs experienced revenue losses during the COVID-19 pandemic, forcing them to cut player salaries and suspend infrastructure projects (Deloitte, 2021). This situation highlights the need for clubs to adopt more careful and strategic transfer policies to ensure financial sustainability.

Player scouting is one of clubs' most crucial tools to discover talented players and strengthen their squads. The scouting process involves analyzing players' performance data to assess their future potential. Performance metrics are used to objectively measure a player's effectiveness and contribution on the field (Mann et al., 2017). These metrics include various statistics such as shooting percentage, pass accuracy, and dribbling success rate.

Scouting platforms play a significant role in providing clubs with player performance data. Platforms like Wyscout and Scout7 offer detailed performance data for thousands of players worldwide. Clubs pay annual subscription fees to access these comprehensive databases (Hudl, 2019). These investments help clubs to make accurate player transfers and minimize financial risks.

Performance metrics are paramount to maximize returns from high-cost transfers. Performance data objective assesse a player's effectiveness and contribution to the team. These data allow coaches and club managers to monitor players' development and make necessary tactical adjustments (Carling et al. et al., 2005). For high-cost transfers to be successful, a player must be talentednd fit into the team's playing style and minimize injury risks. Performance metrics play a critical role in monitoring and managing these factors. Clubs aim to continuously monitor players' performance to maximize returns on their investments (Morgans et al., 2014).

Artificial Intelligence (AI) is increasingly being utilized in the football world. Football teams use AI technologies to analyze player performance, develop tactical strategies, and reduce injury risks. AI-based analysis systems provide coaches valuable insights by examining player movements and team formations in detail during matches. Another significant application of AI in football is player scouting and transfer strategies. AI algorithms analyze the performance data of players worldwide and provide clubs with information on potential transfer targets. This enables clubs to make more informed and data-driven transfer decisions. Additionally, AI technologies are used in injury prediction and prevention systems, making significant contributions to maintaining players' health and performance (Pariath et al., 2018).

This study aims to provide a scientific approach to determining the market values of football players. Through this approach, the goal is to develop an artificial intelligence model that can predict market values based on the future performance of players. This model aims to establish an objective, impartial, and reliable authority in market value determination, thereby contributing to accurate pricing in the football economy, where significant financial transactions occur. Unlike platforms like Transfermarkt, which often rely on subjective inputs such as user votes and expert opinions, this model offers a fully data-driven approach that ensures greater objectivity and precision. The model provides a more comprehensive and detailed valuation by integrating a wide array of performance metrics. Additionally, its ability to forecast future player performance allows clubs to plan their transfer strategies more effectively. In this way, the model will assist clubs in making more informed decisions during transfer processes and support their financial sustainability.

In this study, data collected from 1508 football players playing in the five major European leagues (Spain's La Liga, France's Ligue 1, England's Premier League, Italy's Serie A, and Germany's Bundesliga) and compiled at the end of the 2023-2024 season, from the websites Whoscored and Transfermarkt were analyzed.

A unique dataset was created by collecting 28 different performance metrics, the positions played, and the leagues from WhoScored for each player, while the market value data was gathered from Transfermarkt. Two main processes were performed on this dataset: regression and classification. While predicting the market values of football players based on their performance metrics, the dataset was split into 80% for training and 20% for testing during the regression analyses, whereas k-fold validation with k=5 was employed for the classification tasks. Among seven different regression models, the Random Forest Regression algorithm achieved the highest accuracy, with an R-squared value of 0.90. In the classification tasks, the players were first categorized into four different classes based on their market values, and then their class membership was predicted based on their performance metrics. Among ten different classification algorithms, the highest accuracy was achieved by the CNN algorithm, with a success rate of 97%.

The results indicate that regression and classification models can successfully determine and classify football players' market values. Therefore, it has been demonstrated that a decision support system that can assist football clubs in determining the appropriate transfer fees during the transfer process can be developed using both regression and classification-based artificial intelligence models.

The successful results obtained by the models underscore the importance of performance metrics in determining player market values. These models allow clubs to make more informed decisions about player selection and transfers. Additionally, they contribute to the efficient allocation of transfer expenditures, supporting clubs' financial sustainability.

In recent literature, several studies have explored the prediction of football players' market values using performance metrics and machine learning techniques. The study by Li et al. (2023) explores the use of machine learning models to assess the market values of football players. The research develops two models examining the relationships between players' key characteristics, on-field performance, and salaries. The study by Al-Asadi and Taşdemir (2022) investigates the application of machine learning techniques to predict the market values of football players using FIFA video game data. The research evaluates the effectiveness of various regression models, with the Random Forest algorithm showing the highest accuracy in predicting player market values. The study by Leifheit and Follert (2023) presents a financial valuation approach for football players from a club's perspective, focusing on future payment streams and using the income approach combined with Monte Carlo simulations. The study by Inan and Cavas (2021) develops an artificial neural network (ANN) model to estimate the market values of football players in the Turkish Super League, using performance metrics such as minutes played, goals scored, and passing accuracy. The study by Herm et al. (2014) explores the accuracy and evaluation attributes of an online community (specifically, Transfermarkt.de) in estimating the market values of professional soccer players. The study by Arrul et al. (2022) investigates the application of a neural network model to predict the market values of football players using data from FIFA 19. The study by Behravan and Razavi (2021) introduces a novel machine-learning approach for estimating football players' market values using the FIFA 20 dataset. The study by Kologlu et al. (2018) applies multiple linear regression to estimate the market values of football players in forward positions, using physical and performance factors from the 2017-2018 season. Lee et al. (2022) study proposes an optimized LightGBM model using Bayesian hyperparameter optimization to predict football players' market values. Aydemir et al. (2022) study proposes a machine-learning ensemble approach to predict football players' transfer values. The study by Franceschi et al. provides a systematic review of empirical research aimed at identifying the factors influencing football players' market valuations (Franceschi, 2024). The studies encountered in the literature on predicting football players' market values using artificial intelligence techniques have been listed above. A detailed comparison of each study with our work is provided in the discussions section.

METHODS

This section describes the dataset employed in the study, the algorithms used, and the metrics utilized to evaluate the data derived from these algorithms.

Dataset

The dataset used in this study contains performance data and market values for 1508 football players playing in the top five European leagues, with 28 different performance metrics included. Additionally, the dataset includes information on the players' positions and the leagues they play. The performance data, league information, and positions of the players were individually collected from the Whoscored platform. Subsequently, the market value of each player was obtained from the Transfermarkt platform. (Whoscored, 2024; Transfermarkt, 2024). This approach resulted in a comprehensive dataset on players in Europe's major leagues. The independent and dependent variables included in the dataset are presented in Table 1. The minimum and maximum value ranges for each variable are also shown in the same table.

Table 1
The Features Within the Dataset and Their Value Ranges

Variables Type	Features	Min – Max Range
Independent Variables	Age	16-40
	Minutes played	173-3060
	The number of matches started in the first 11	0-34
	Number of matches in which he was substituted	0-24
	Total goals	0-33
	Total assists	0-12
	Yellow card	0-16
	Red card	0-3
	Pass success percentage	46.1-95.5
	Aerial duels won per game	0-6.8
	Man of the match	0-9
	Shots per game	0-4.5
	Dribbles per game	0-3.6
	Fouled per game	0-2.9
	Offsides per game	0-1.5
	Dispossessed per game	0-2.9
	Bad control per game	0-3.8
	Tackles per game	0-4.6
	Interceptions per game	0-2.4
	Fouls per game	0-2.7
	Offside won per game	0-1.7
	Clearances per game	0-6
	Outfielder block per game	0-1.6
	Own goals	0-2
Independent Variables	Key passes per game	0-3.7
	Passes per game	3.1-106
	Crosses per game	0-2.9
	Long balls per game	0-15.2
	Through balls per game	0-0.6
	League	Categorical
	Position	Categorical
Dependent Variables	Market Value (€)	0-180000000

Scaling The Features in the Dataset

Min-Max Scaling is a widely used method in data normalization. This technique normalizes data by rescaling each value in a dataset to a specified range. Min-Max Scaling is particularly beneficial in machine learning models, where it helps make data points on different scales comparable. The method rescales each data point based on the minimum and maximum values in the dataset, so all data falls within a specified range (typically [0, 1]). Mathematically, for a data point x, Min-Max Scaling is expressed as:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

In this formula, x' represents the scaled value, x is the original value, $x_{\min}$ is the minimum value in the dataset, and $x_{\max}$ is the maximum value. This method enables the effective processing of different variables, particularly those measured in different units or scales, by machine learning algorithms (Jain et al., 2000).

Min-Max Scaling ensures that all data in the dataset falls within a specified range, making it particularly effective in gradient-based optimization algorithms and rule-based learning methods (Han et al., 2011). However, a major limitation of Min-Max Scaling is that if new values fall outside the range of the original minimum and maximum values, the scaled values may fall outside the [0, 1] range. Therefore, Min-Max Scaling works best when the same minimum and maximum values are used for both the training and test datasets.

Model Selection

In this study, the regression analyses were performed using the following algorithms: Adaboost, Decision Tree, Gradient Boosting, K Nearest Neighbors, Random Forest, Ridge Regression, and Support Vector Machine. For classification tasks, various algorithms were tested, including RandomForestClassifier, GaussianNB, SVC, KNN, GradientBoostingClassifier, XGBClassifier, AdaBoostClassifier, LogisticRegression, MLPClassifier, and CNN, with extensive hyperparameter tuning via GridSearchCV. This section explains the Random Forest and Gradient Boosting algorithms, which achieved the highest performance in the regression analysis, the CNN algorithm, which yielded the best results in the classification tasks.

Random Forest

Random Forest is another ensemble learning method widely used for classification and regression tasks. Unlike Gradient Boosting, which builds trees sequentially, Random Forest constructs many decision trees simultaneously during training and aggregates their results.

This technique helps reduce overfitting and improves the model's generalization ability. The prediction for a new input x in a Random Forest is the average of the predictions from all individual trees:

$$\tilde{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x)$$

Where $T_i(x)$ represents the prediction from the i-th tree in the forest, and N is the total number of trees. Each tree in the Random Forest is built using a bootstrap sample of the data, and at each split, a random subset of features is selected. This process introduces randomness into the model, which helps variance and avoid overfitting (Breiman, 2001). The construction of each tree involves recursively splitting the data at points that maximize the reduction in a loss function, such as mean squared error for regression tasks.

This study applied the Random Forest algorithm to the same dataset of football players' performance metrics. The model demonstrated strong predictive power, making it a valuable tool for estimating market values. The inherent capability of Random Forest to handle many features and capture non-linear relationships contributed to its success in this context. The aggregated results from multiple trees provided a robust and stable prediction, enhancing the reliability of the model (Liaw & Wiener, 2002).

Gradient Boosting

Gradient Boosting is a powerful machine learning technique that builds an ensemble of weak learners, typically decision trees, to create a strong predictive model. The core idea of Gradient Boosting is to add trees to the model, each correcting the errors of the previous trees. This process can be mathematically expressed as follows: Given a dataset (X, y), where X represents the input features and y the target variable, the goal is to minimize a loss function $L(y, \tilde{y})$ over the predictions $\tilde{y}$.

Initially, the model starts with a simple prediction, such as the mean of y:

$$\tilde{y}_0 = mean(y)$$

In each subsequent iteration m, a new tree $h_{\mathrm{m}}(X)$ is trained to fit the negative gradient of the loss function with respect to the current model prediction:

$$T_{i,m} = -[\frac{\partial L(yi, \tilde{y}i)}{\partial \tilde{y}_i}]_{\tilde{y}_i - \tilde{y}_{i,m-1}}$$

The model is then updated as follows:

$$\tilde{y}_m = \tilde{y}_{m-1} + \lambda h_m(X)$$

Here, λ is the learning rate that controls the contribution of each tree. This iterative process continues until the model's performance no longer improves significantly on a validation set. Gradient Boosting has been shown to be highly effective in various predictive tasks due to its ability to capture complex patterns in the data (Friedman, 2001).

Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a class of deep learning algorithms particularly well-suited for tasks involving image and spatial data analysis, but they have also been successfully applied to various other classification problems. A CNN consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers apply a series of filters (or kernels) that slide across the input data, detecting local patterns and features. Mathematically, the convolution operation performed by the filters can be expressed as:

$$S(i,j) = (I * K)(i,j) \sum_{m} \sum_{n} I(m,n).K(i-m,j-n)$$

Where $S(i,j)$ is the output feature map, I is the input matrix, and K is the convolutional kernel. This operation allows the CNN to automatically and hierarchically learn spatial hierarchies of features from the input data.

In classification tasks, the output from the convolutional and pooling layers is typically passed through fully connected layers, which interpret the learned features and make predictions. The final layer of the CNN usually uses a softmax function to produce a probability distribution over the possible classes. The CNN's parameters, including the filter weights, are learned during training through backpropagation and gradient descent optimization. This architecture enables CNNs to achieve high accuracy in various classification tasks, making them a powerful tool for machine learning (LeCun et al., 1998; Goodfellow et al., 2016).

Evaluation of the models

In this study, the performance of the regression models was evaluated using the R^2, MSE (Mean Squared Error), and RMSE (Root Mean Squared Error) metrics, while the performance of the classification models was assessed using the accuracy, precision, recall, F1-score, and support metrics.

R-Squared (R^2)

R-squared is a statistical measure represents the proportion of the variance for a dependent variable explained by an independent variable or variables in a regression model. It is often used to evaluate the goodness of fit of a regression model. The value of R^2 ranges from 0 to 1, where 0 indicates that the model explains none of the variability of the response data around its mean, and 1 indicates that the model explains all the variability. The R^2 value can be calculated using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \tilde{y})^2}$$

In this formula, y_i represents the observed values, $\tilde{y}_i$ represents the predicted values, and $\tilde{y}$ is the mean of the observed values (Draper & Smith, 1998). R^2 is a key metric because it provides an indication of how well the model's predictions match the actual data.

Mean Squared Error (MSE)

Mean Squared Error (MSE) is a common measure of the quality of an estimator—it is always non-negative, and values closer to zero are better. MSE is the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. The formula for MSE is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

In the formula above, n is the number of observations, y_i is the actual value, and $\tilde{y}_i$ is the predicted value (Montgomery et al., 2012). MSE is particularly useful because it penalizes larger errors more severely than smaller ones, making it a sensitive measure of prediction accuracy.

Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is the square root of the Mean Squared Error. It is used to measure the differences between values predicted by a model and the values observed. RMSE is a good measure of how accurately the model predicts the response and is the standard deviation of the prediction errors (residuals). The formula for RMSE is:

$$MSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}$$

In the formula above, n is the number of observations, y_i is the actual value, and $i\tilde{y}_i$ is the predicted value (Chai & Draxler, 2014). RMSE is widely used because it gives an easily interpretable measure of model accuracy in the same units as the data, and it is more sensitive to outliers than MSE.

Accuracy

Accuracy is one of the most used metrics for evaluating the performance of a classification model. It represents the ratio of correctly predicted instances to the total number of instances. Mathematically, the accuracy metric is defined as:

$$Accuracy: \frac{TP + TN}{TP + TN + FP + FN}$$

In this formula, TP (True Positives) denotes the instances correctly classified as positive, TN (True Negatives) represents the instances correctly classified as negative, FP (False Positives) refers to the instances incorrectly classified as positive, and FN (False Negatives) represents the instances incorrectly classified as negative. Accuracy provides a general measure of how well the classification model performs overall, but it can be misleading in cases of class imbalance (Fawcett, 2006; Sokolova & Lapalme, 2009).

Precision

Precision is a performance metric that indicates how many instances predicted by a classification model are positive. Precision is critical when assessing a model's ability to minimize false positives. Mathematically, the precision metric is defined as:

$$Precision = \frac{TP}{TP + FP}$$

In this formula, TP (True Positives) represents the instances correctly classified as positive, while FP (False Positives) refers to the instances incorrectly classified as positive. Precision plays a critical role in measuring the reliability of a classification model's otmispositic predictions, especially in imbalanced datasets (Powers, 2011; Sokolova & Lapalme, 2009).

Recall

Recall is a performance metric that indicates how effectively a classification model can identify positive instances. It is particularly important for assessing the model's ability to minimize the number of false negatives. Mathematically, the recall metric is defined as:

$$Recall = \frac{TP}{TP + FN}$$

In this formula, TP (True Positives) represents the instances correctly classified as positive, while FN (False Negatives) refers to the instances incorrectly classified as negative, meaning the positive instances that the model missed. Recall plays a critical role, especially in imbalanced datasets, for measuring how well the model can capture all instances of the positive class (Sokolova & Lapalme, 2009; Manning et al., 2008).

F1-Score

The F1-score is a performance metric defined as the harmonic mean of precision and recall. It balances precision and recall, ensuring that both metrics are considered. This is particularly important in imbalanced datasets or when the model needs to balance false positives and negatives. Mathematically, the F1-score is expressed as:

$$F1 - score = 2 * \frac{precision * recall}{precision + recall}$$

In this formula, precision represents the proportion of true positives among the instances predicted as positive, while recall indicates how well the model identifies all positive instances. The F1-score is recognized as a critical metric for balancing the disparity between precision and recall and for evaluating the model's overall performance (Powers, 2011; Sokolova & Lapalme, 2009).

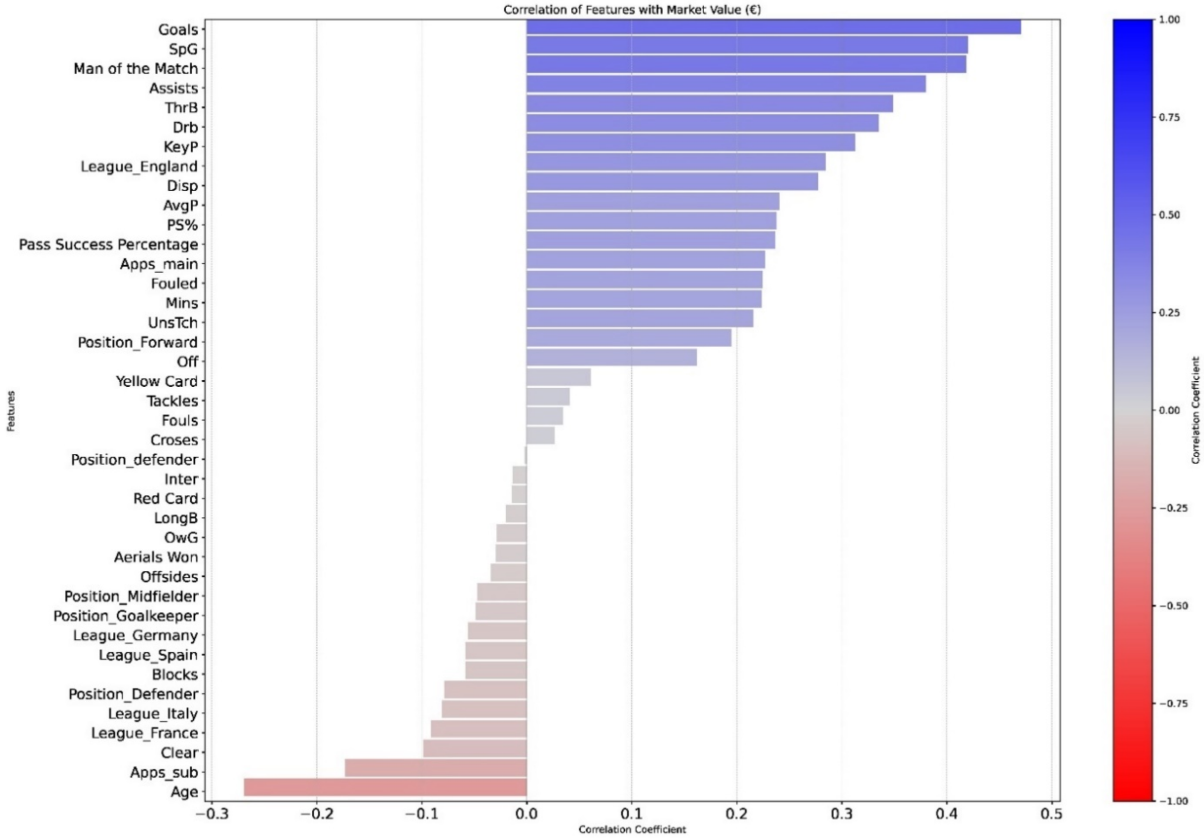
RESULTS

In this study, two different methods, regression and classification, were applied to the original dataset collected. In the regression analysis, the dependent variable, Market Value, was predicted using other independent attributes. In the classification analysis, the football players in the dataset were categorized into different classes based on their market values. Subsequently, the cost classes of the players were predicted using the independent variables.

Regression Analysis

In this study, seven different regression models (Adaboost, Decision Tree, Gradient Boosting, K Nearest Neighbors, Random Forest, Ridge Regression, and Support Vector Machine) were initially applied to the raw dataset. Due to the unsatisfactory accuracy of the results from these procedures, the dataset was first adjusted using the standard scaling method. Subsequently, a correlation analysis was conducted to identify and remove features unrelated with the dependent variable. Figure 1 shows features with the highest positive and negative correlations with the dependent variable, Market Value.

Figure 1
Top 10 Positive and Negative Correlations with Market Value



According to the correlation analysis performed, the independent variables with the highest positive correlations with the dependent variable, market value, are as follows: Total goals (Goals), shots per game (SpG), man of the match, assists, through balls per game (ThrB), sribbles per game (Drb), key passes per game (KeyP), being in the Premier League (League_England), dispossessed per game (Disp), AvgP, passes per game (PS%), pass success percentage, top 11 (Apss_main), fouled, playing time (Mins), turnover per game (UnsTch), forward (Position_Forward), off, age, number of matches in which he was substituted (Apps_sub), clearances per game (Clear).

The independent variables with the highest negative correlations with the dependent variable, market value, are as follows: Yellow card, tackles, fouls, interceptions per game (Inter), red card, long balls per game (LongB), own goals (OwG), aerial duels won per game (Aerials Won), offsides per game (Offsides), midfielder (Position_Midfielder), goalkeeper (Position_Goalkeeper), being in the Germany League (League_Germany), Being in the Spain League (League_Spain), outfielder block per game (Blocks), defender (Position_Defender), being in the Italy League (League_Italy), being in the France League (League_France).

Following the correlation analysis, the attributes with the highest positive and negative correlations with the dependent variable were retained in the dataset, while the remaining attributes were removed. With the refined dataset, the dependent variable, Market Value, was predicted using the following algorithms: Adaboost, Decision Tree, Gradient Boosting, K Nearest Neighbors, Random Forest, Ridge Regression, and Support Vector Machine. The Random Forest and Gradient Boosting algorithms obtained the most successful results. Therefore, the data and results pertaining to these algorithms are presented. In this study, we opted not to use traditional reference groups for the League and Position variables due to their overall weak explanatory power. Only League_England and Position_Forward demonstrated a meaningful correlation with the target variable, while other leagues and positions showed weaker or insignificant effects on market value.

The hyperparameters used while running the Random Forest and Gradient Boosting algorithms were fine-tuned to maximize the performance of each algorithm. The most successful hyperparameter values obtained are shown in Table 2.

Table 2
The Best Hyperparameter Values Obtained for the Machine Learning Algorithms Used

Algorithms	Trained Hypermeters	Best Hypermeters Values
Gradient Boosting	'n_estimators' : [50, 75, 100], 'max_depth': [2, 4, 8], 'min_samples_split' : [1,2,4], 'min_samples_leaf' : [1,2 ,8], 'learning_rate' : [0.05, 0.1 , 0.5, 1]	learning_rate': 0.05, 'max_depth': 4, 'min_samples_leaf': 8, 'min_samples_split': 4, 'n_estimators': 100
Random Forest	'n_estimators' : [50,100,200], 'max_depth': [2, 4, 8, 16], 'min_samples_split' : [2,3,4], 'min_samples_leaf' : [2,3,4,8]	max_depth': 16, 'min_samples_leaf': 4, 'min_samples_split': 4, 'n_estimators': 50

Table 3 presents the training and testing graphs, prediction error distributions, and validation graphs for the Random Forest and Gradient Boosting algorithms, which were trained to predict the market values of football players in the top five European leagues using machine learning. A linear plot is expected in the training and testing graphs for successful training. Although the Random Forest algorithm's test and training graphs do not show a clear linear line, it can be said that there is no significant deviation. In the error distribution graphs, distributions with minimal variation are expected. When examining the error distribution graphs for the Gradient Boosting and Random Forest algorithms, it is observed that the errors are concentrated around zero. In the prediction and accuracy graphs, the blue lines represent the actual values, while the orange lines represent the values predicted by the model. In this graph, the desired outcome is the overlap of these two different colored lines. The graph obtained from the Gradient Boosting algorithm ishows that the blue and orange lines overlap.

Table 3

Various Performance Graphs Obtained From Random Forest and Gradient Boosting Algorithms

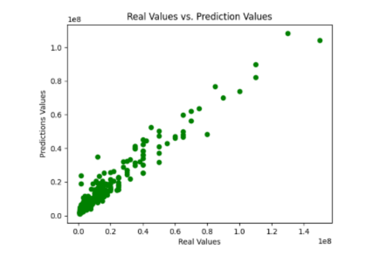
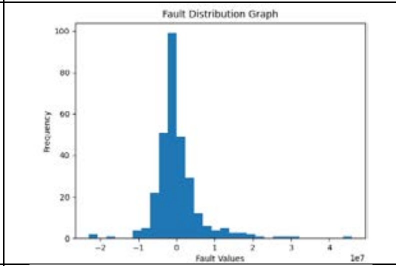
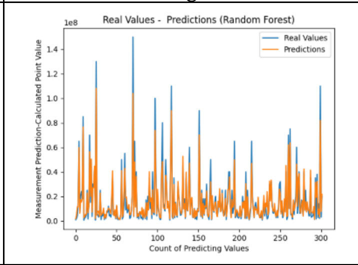
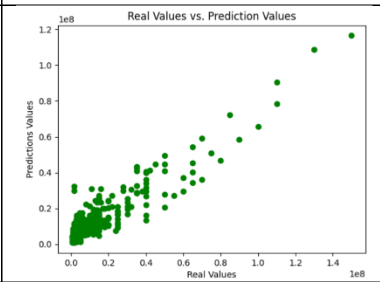
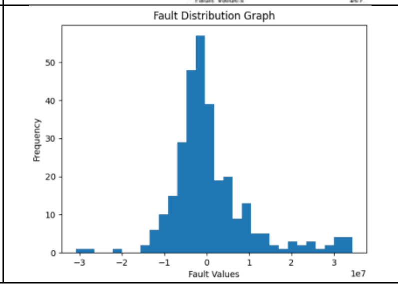
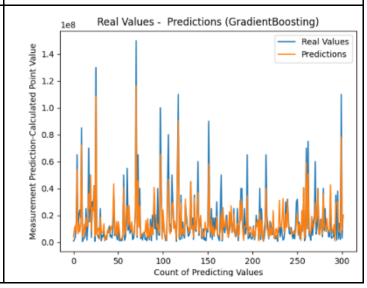
Alg.	Real Values & Prediction	Fault Distribution	Predicting Values
Random Forest			
Gradient Boosting			

Table 4 shows the metric values obtained from Gradient boosting and random forest algorithms. The metrics obtained from the Gradient Boosting and Random Forest regression models provide significant insights into the models' performance and accuracy in predicting football players' market values.

Table 4

Various Performance Metrics Obtained From the Algorithms Used

Algorithms	R-Squared	MSE	RMSE
Random Forest	0.9040082175222088	43761946912597.55	6615281.317721684
Gradient Boosting	0.80772276818793	87657774383008.19	9362573.064228026

Classification Analysis

In this study, classification processes were conducted following the regression procedures on the original dataset. The subsequent sections provide information on the reasons for performing classification, how the classes were created, and the metric values obtained from these procedures.

The market value has been converted to categorical named “Low”, “Lower-Mid”, “Upper-Mid” and “High”. Converting a continuous target into a categorical one can significantly influence predictive models' effectiveness and applicability. Predictive models typically predict a quantity when dealing with continuous variables, thus framing the problem as a regression. Conversely, if the target variable is categorical, the task transforms into a classification aimed at predicting discrete labels. In predictive modeling, the characterization of the target variable crucially defines the analytical strategy—whether one might pursue regression or classification methodologies (Gareth et al., 2013). The conversion of continuous variables into categorical variables is driven by several considerations that enhance the interpretability and applicability of the analytical outcomes:

- *Interpretability:* Categorical outcomes simplify interpreting the model's results, making the predictions more intuitive and actionable (Hastie et al., 2009).
- *Handling Non-linear Relationships:* Classification algorithms can effectively manage complex, non-linear relationships that linear regression might fail to capture, thus potentially improving model performance (Breiman, 2001).
- *Dealing with Skewed Data:* Continuous data that are heavily skewed may adversely affect model performance, especially if certain statistical assumptions (e.g., normality) are violated. Categorization can mitigate such issues (George et al., 2005).
- *Robustness and Performance:* Models built on categorized data are often more robust and simpler to validate across different subpopulations, thus enhancing the model's generalizability and utility (Breiman, 2001).

The transformation from continuous to categorical variables should adhere to sound methodological principles. The categorization process must ensure that critical information is not obscured and the resulting categories are substantively meaningful (Agresti, 2012).

In feature selection, correlation and ANOVA tests have been widely used in machine learning. The correlation between a feature and the target value can vary widely depending on the nature of the data and the specific problem. There isn't a specific threshold for a "good" correlation value that applies universally, but understanding the correlation's strength can help us select feature and understand the underlying relationships in the data. With correlation, +1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear correlation. It can be said that if an absolute value of correlation is near 1, then the relationship between the feature and the target value is high, but this also raises concerns about potential overfitting if the model is overly reliant on them.

On the other hand, when it comes to ANOVA tests, the F-values and p-values help identify the features that have a statistically significant relationship with the target variable. Of course, there are neither "good" values for F-values or p-values. However, these values can be very helpful in understanding the data characteristics and the desired level of confidence in the results.

The F-value represents the variance ratio between the groups (between-group variability) to the variance within the groups (within-group variability). A higher F-value indicates that the group means are significantly different, suggesting a stronger relationship between the feature and the target variable. Thus, it can be said that higher F-values are better. The p-value indicates the probability that the observed data could have occurred by chance under the null hypothesis, which in the context of ANOVA, there are no differences among group means. A low p-value suggests that the observed data are unlikely under the null hypothesis and that there is a statistically significant effect. Typically, a p-value less than 0.05 is considered statistically significant, meaning there is less than a 5% probability that the difference among the group means occurred by chance. However, this threshold can be adjusted if we want to be more confident and reduce the chance of false positives, adopting a more stringent level like 0.01.

Table 5 displays the F-Values and P-Values from the ANOVA tests and the correlation values between features and the target variable. The table is sorted based on the F-Value. A p-value threshold of 0.01 was chosen, along with high F-values, to provide a more stringent criterion for statistical significance. This reduces the likelihood of false positives, enhances confidence in the results, and aligns with rigorous scientific standards. Additionally, a correlation threshold 0.20 was determined to ensure strong relationships between features and the target variable. Consequently, the features 'Goals', 'League_England', 'Age', 'ThrB', 'Man of the Match', 'SpG', 'Assists', 'Drb', 'KeyP', 'AvgP', 'Disp', 'PS%', 'Pass Success Percentage', 'Apps_main', 'Mins', 'Fouled' and 'UnsTch' were selected for our model.

Table 5
F-Values and P-Values From the ANOVA Tests

Feature	F-Value	P-Value	Correlation
Goals	6.921477	1.016068e-56	0.470555
League_England	6.203801	3.669312e-49	0.284496
Age	5.659184	2.187875e-43	-0.269271
ThrB	5.258585	3.975027e-39	0.348716
Man of the Match	5.230734	7.862186e-39	0.418369
SpG	5.198296	1.739901e-38	0.420393
Assists	4.270797	1.172761e-28	0.380256
Drb	3.327598	6.652129e-19	0.335281
KeyP	3.251458	3.875351e-18	0.312894
AvgP	2.815660	7.337121e-14	0.240314
Disp	2.750623	3.058217e-13	0.277440

Table 5 *(Continued)*

Feature	F-Value	P-Value	Correlation
PS%	2.705459	8.177499e-13	0.237691
Pass Success Percentage	2.687372	1.210252e-12	0.236792
Apps_main	2.450002	1.860794e-10	0.227125
Mins	2.379942	7.874275e-10	0.223510
Fouled	2.283613	5.513421e-09	0.224452
Feature	F-Value	P-Value	Correlation
League_Spain	2.067382	3.631018e-07	-0.058138
UnsTch	1.880720	1.058149e-05	0.215552
Position_Forward	1.853433	1.694458e-05	0.194655
Off	1.768668	7.026165e-05	0.162074
League_Italy	1.636212	5.666960e-04	-0.080680
Apps_sub	1.629456	6.273308e-04	-0.173085
Tackles	1.401693	1.404777e-02	0.040936
League_Germany	1.369985	2.050243e-02	-0.056099
Yellow Card	1.354213	2.460770e-02	0.060769
League_France	1.337613	2.969716e-02	-0.091227
Position_Defender	1.223311	9.606589e-02	-0.078811
Fouls	1.103788	2.563111e-01	0.034875
Inter	1.064114	3.342108e-01	-0.013796
Position_Midfielder	1.034925	3.982358e-01	-0.047409
Blocks	0.993588	4.957139e-01	-0.058723
Red Card	0.941609	6.219149e-01	-0.014327
OwG	0.931666	6.455674e-01	-0.028492
Clear	0.926909	6.567548e-01	-0.098732
Offsides	0.846132	8.250200e-01	-0.034334
Croses	0.789850	9.076117e-01	0.026532
Aerials Won	0.729231	9.616859e-01	-0.029563
LongB	0.702325	9.759380e-01	-0.020089
Position_Goalkeeper	0.671243	9.868230e-01	-0.048497

The classification model has been tested using various algorithms, including RandomForestClassifier, GaussianNB, SVC, KNN, GradientBoostingClassifier, XGBClassifier, AdaBoostClassifier, LogisticRegression, MLPClassifier, and CNN, with extensive hyperparameter tuning via GridSearchCV. The best performance was achieved with the CNN model, which attained an accuracy of 97%. Other algorithms did not surpass an accuracy of 80%. Table 6 outlines the layer architecture of the CNN model. Sixteen features were selected for the model, necessitating multiple convolutional layers to capture complex hierarchical features. BatchNormalization layers stabilize the model and accelerate training, MaxPooling layers reduce spatial dimensions and computational load, and Dropout layers prevent overfitting. These architectural choices collectively create a powerful and efficient model that accurately classifyies data with the selected 16 features. Table 7 displays the metric values for the classification process performed using the CNN model. Figure 2 shows the confusion matrix for the classification process performed using the CNN model.

Table 6
Layer Architecture and Parameters of the CNN Model for Classification Task

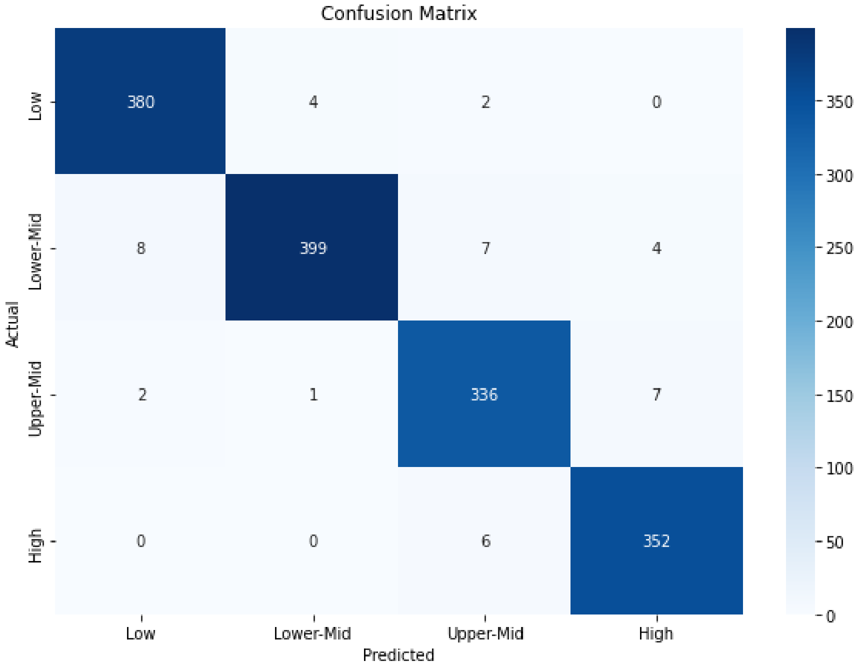
Layer (type)	Output Shape	Param #
conv1d_3 (Conv1D)	(None, 17, 64)	192
batch_normalization_3	(Batch (None, 17, 64))	256
max_pooling1d_3	(MaxPooling1 (None, 8, 64)	0
conv1d_4	(Conv1D) (None, 8, 64)	8256
batch_normalization_4	(Batch (None, 8, 64)	256
max_pooling1d_4	(MaxPooling1 (None, 4, 64)	0
conv1d_5	(Conv1D) (None, 4, 32)	4128
batch_normalization_5	(Batch (None, 4, 32)	128
max_pooling1d_5	(MaxPooling1 (None, 2, 32)	0
flatten_1	(Flatten) (None, 64)	0
dense_3	(Dense) (None, 256)	16640
dropout_2	(Dropout) (None, 256)	0
dense_4	(Dense) (None, 128)	32896
dropout_3	(Dropout) (None, 128)	0
dense_5	(Dense) (None, 4)	516
Total params: 63.288 **Trainable params: 62.948** **Non-trainable params: 320**		

Table 7
The Metrics for the CNN Model

	Precision	Recall	F1- Score	Support
0	0.97	0.98	0.98	386
1	0.99	0.95	0.97	418
2	0.96	0.97	0.96	346
3	0.97	0.98	0.98	358
Accuracy			0.97	1508
Macro Avg.	0.97	0.97	0.97	1508
Weighted Avg.	0.97	0.97	0.97	1508

Note. Test Accuracy: 0.9728116989135742

Figure 2
The Confusion Matrix of the Proposed CNN Model



DISCUSSION

This study has demonstrated the significant role of performance metrics in estimating the market values of football players in Europe's top five leagues.

In this study, the highest R-squared value for determining football player market value using regression analysis was 0.90 with the Random Forest algorithm. When examining the performance of the Random Forest model, the Mean Squared Error (MSE) was found to be 43761946912597.55. Although this value appears to be extremely high, it is understandable given that the dataset is in Euros and has a wide range of player values. Football players' market values can vary from a few hundred thousand Euros to hundreds of millions of Euros. This broad range leads to substantial prediction errors. The Root Mean Squared Error (RMSE) was calculated to be 6615281.31. Since RMSE is the square root of MSE, it represents the error in the same unit, indicating that the model's predictions deviate from the actual market values by an average of approximately 6.9 million Euros. This error margin is acceptable considering the natural variability in football players' market values. The model's R-squared (R2) value was determined to be 0.90, indicating that the model explains 90% of the variance in the dependent variable (market value), which signifies a strong performance.

The obtained R^2 value indicates that the overall performance and accuracy of the models provide reliable results in predicting football players' market values. Despite the high MSE and RMSE values, which are expected due to the nature of the dataset and the wide range of market values, the strong R2 value demonstrates the models' effectiveness in capturing the underlying patterns in the data.

Random Forest emerged as the most successful algorithm in regression analysis due to its ensemble learning approach, which combines multiple decision trees to improve predictive accuracy and control overfitting. Each decision tree in the Random Forest is trained on a random subset of the data, allowing the model to capture a wide range of patterns and relationships within the dataset. This method is particularly effective in handling complex, non-linear interactions between features, making it more robust against noise in the data. In contrast, other algorithms like Ridge Regression and Support Vector Machines struggled to achieve similar accuracy due to their linear assumptions and sensitivity to outliers, which limited their ability to model the intricate patterns in the data.

In this study, the highest accuracy rate for determining football player market value using a classification process was 97% with the CNN algorithm. The metrics obtained from the classification process using the CNN model indicate a very high performance. The average precision, recall, and F1-score of 0.96 to 0.99 were calculated based on the class. These values demonstrate the model's high accuracy in positive predictions (precision) and its effectiveness in positive classification (recall).

An F1-score of 0.96 to 0.98 show a good balance between precision and recall, indicating strong overall performance of the model. The average support value of 1509 signifies that there are enough data points in each class, ensuring the reliability of the results. Additionally, a test accuracy of 97% suggests that the model has a high general accuracy and can correctly classify most of the examples in the dataset. These metrics indicate that the CNN model performs effectively and reliably in the classification task, and it also highlights the model's high generalizability.

In the classification tasks, the Convolutional Neural Network (CNN) outperformed other models due to its ability to automatically and hierarchically extract features from the input data, mpartinly when dealing with spatial or grid-like structures. CNNs are designed to capture local patterns through convolutional layers, which is especially advantageous when the data has underlying spatial dependencies or complex feature interactions. This allows the CNN to effectively differentiate between classes, even in the presence of noisy or irrelevant data. On the other hand, traditional models like Logistic Regression or even other machine learning algorithms such as RandomForestClassifier and SVC struggled because they either could not automatically learn feature representations or were less effective in capturing the complex relationships within the data.

In our analysis, we observed negative correlations between spcifcertaic defensive metrics, such as interceptions per game, aerial duels won per game, clearances per game, and players' market values. Although these metrics are typically seen as indicators of strong defensive performance, their negative correlation with market value can be explained by several factors upon closer examination.

First, it is important to consider the positional differences in football. Players who excel in these defensive statistics are typically defenders and, in some cases, defensive midfielders. Historically, these positions tend to have lower market values compared to attacking players, who contribute directly to goal-scoring opportunities. For example, forwards and attacking midfielders who accumulate goals, assists, and key passes are generally valued more highly by clubs due to their perceived direct impact on match outcomes. This disparity in market value between positions can explain why metrics strongly associated with defenders, such as clearances and aerial duels, show negative correlations with market value. The underlying logic is not that these players are less valuable but rather that their market value is shaped by their positional role and perceived importance in modern football.

Secondly, league-level differences also play a role. The Premier League, for instance, is known for its more physically demanding style of play, where defensive contributions such as aerial duels and clearances might be more prominent compared to other leagues like La Liga or Serie A, which may emphasize technical skills and ball control. The overall market value of players in the Premier League is typically higher due to the financial strength of clubs in this league. However, as these defensive metrics are less valued in attacking players, it may result in a negative correlation between these statistics and market value, particularly when aggregated across different leagues. In conclusion, while these defensive metrics are undoubtedly crucial to a team’s overall success, their impact on individual market values is mediated by positional roles and league-specific dynamics. Future research could benefit from applying normalization techniques to better capture the true relationship between player performance and market value across various contexts.

This study's primary aim is to estimate football players' global market values based on their performance data. Accordingly, the objective is to analyze the impact of performance metrics on individual player valuations from a global perspective. Our study focuses not on examining differences between leagues or positions, but rather on developing a general market valuation model applicable to all football players.

While normalization by league and position could offer a more nuanced analysis, it would shift the study's focus towards localized market dynamics, which falls outside the scope of this study. The central goal of this research is to provide a global assessment, offering a broader perspective on player valuation in the world of football. Normalization based on league and position would be more suited for studies analyzing local differences. However, this approach does not align with the purpose of the porposed study, which aims to estimate market values from a universal perspective.

Furthermore, estimating global market values based on performance data aims to offer clubs a general guide for their transfer strategies. Emphasizing league and position-based differences would require specific models for each club or league, which would narrow the scope of our study and weaken its general applicability. Therefore, we believe that normalization according to league and position averages falls outside the scope and purpose of this research.

The results of the regression and classification models offer insights into various aspects, such as football performance, the transfer market, and football club economics, providing a foundation for practical applications in club management and strategic transfers.

Performance metrics provide an objective way to measure players' impact on the field. In this study, metrics such as the number of goals, shots per game, assists, man of the match and age were found to have high correlations with players' market values. These findings underscore the importance of adopting a more analytical approach in evaluating player performance. Coaches and technical directors can use these metrics to better analyze players' contributions on the field and make strategic decisions.

The findings of this study can directly influence decision-making processes in the football transfer market. Player transfers are critical for clubs' success, and mistakes in these processes can lead to significant financial losses. Our study has shown that regression and classification models based on performance metrics can accurately predict player market values. This encourages clubs to adopt more precise and data-driven approaches in their transfer decisions.

Football economics is vital for clubs' financial health and sustainability. Transfer expenditures constitute a large portion of clubs' budgets, and poor transfer policies can lead to financial crises. Our study has demonstrated that performance data can be effectively used to estimate players' market values. This can help clubs make more informed decisions during transfers and utilize their financial resources more efficiently.

Clubs' economic sustainability is critical for maintaining their long-term success. Since many clubs experienced revenue losses during the COVID-19 pandemic, cost-effective transfer strategies have become even more important. This study provides valuable insights that can help clubs optimize their transfer expenditures and reduce financial risks. Prediction models based on performance metrics can enable clubs to make more strategic player selections, supporting their financial sustainability.

Various studies have employed machine learning and statistical modeling techniques to estimate the market values of football players. While some of these studies share similarities in methods and findings with our work, distinct differences set our research apart.

Li et al. (2023) focused on using machine learning models to predict football players' market values, primarily using player salaries as proxies. In contrast, our research directly estimates market values by incorporating additional variables such as position and league information into a broader dataset and by comparing the performance of different machine learning models to provide a more comprehensive approach. This distinction highlights our contribution to achieving more accurate predictions.

Al-Asadi and Tasdemir (2022) utilized FIFA video game data to estimate player market values. Our approach differs by using real-world performance data and including additional variables like league and position. Furthermore, we compare multiple machine learning models and optimize them to enhance predictive accuracy, making our study more applicable to real-world scenarios.

Leifheit and Follert (2023) introduced a financial valuation model that focuses on players' subjective and future-oriented value based on payment streams. Our study diverges by directly predicting market values using a variety of machine learning models, incorporating real-world performance data such as position and league, and offering a more practical and data-driven model for market value prediction.

Inan and Cavas (2021) applied an artificial neural network (ANN) to estimate market values in the Turkish Super League. While their study provides valuable insights into ANN application, our research is distinguished using a broader range of machine learning models and additional variables like league and position data. We also emphasize model optimization to enhance accuracy, providing a more comprehensive approach.

Herm et al. (2014) examined crowd-sourced evaluations of soccer players’ market values, highlighting external popularity metrics and talent-related factors. Our research, however, leverages advanced machine learning models and systematically incorporates performance metrics and additional variables, resulting in a more objective and data-driven prediction model.

Arrul et al. (2022) focused on a neural network model using FIFA 19 data. In contrast, our research incorporates real-world performance data and additional variables like league and position. We also compare and optimize multiple machine learning models, offering a more comprehensive approach to player valuation.

Behravan and Razavi (2021) emphasized a hybrid machine learning model using APSO-clustering combined with PSO-SVR based on video game data. Our study expands on this by using real-world data and incorporating additional variables. We apply multiple machine learning models in a real-world context, optimizing them to enhance accuracy and practical application in professional football.

Koloğlu et al. (2018) employed multiple linear regression to assess market values, particularly for forward positions. While their study provides insights into influential factors, our research differs by using a broader range of machine learning models, including ensemble methods, and incorporating variables like league and position data for a more dynamic approach.

Lee et al. (2022) introduced an optimized LightGBM model with Bayesian hyperparameter optimization. While their study focuses on optimizing a single model, our research compares multiple models and incorporates additional variables like league and position data, providing a more versatile approach to predicting market values.

Aydemir et al. (2022) used a machine learning ensemble approach with diverse data sources to predict transfer values. Our research extends this by incorporating additional variables like league and position data and comparing multiple models optimized for real-world scenarios, offering a more adaptable and robust framework.

Franceschi et al. (2024) systematically reviewed the factors affecting market valuations, focusing on a general framework. Our research, however, employs advanced machine learning models, incorporates specific variables like position and league data, and compares a greater variety of models, setting our work apart as a significant contribution to more accurate predictions. Compared to these studies, the original contributions of our work include:

- *Multiple Model Comparison:* We compared various machine learning models to identify the best-performing ones in predicting market values.
- *Use of Ensemble Methods:* We developed more robust predictive models by reducing noise and improving accuracy through ensemble methods.
- *Hyperparameter Optimization:* Optimizing the models' hyperparameters significantly enhanced their predictive accuracy.

These aspects demonstrate that our study makes a significant contribution to the literature, representing an important step toward more accurate estimation of football players' market values.

One of study's major limitations is the data collection challenge. Specifically, no existing platform provides all the necessary information, making the process complex. The authors meticulously collected data for approximately 1,500 players, including performance, league, and position information. Despite being time-consuming and labor-intensive, this process was carried out with great care to ensure reliability and validity.

Based on our findings, we recommend that future research focus on expanding the dataset to include a broader range of leagues, seasons, and player groups. Utilizing larger and more diverse datasets could improve generalizability and provide a more comprehensive understanding of the factors influencing player market values.

In conclusion, this study highlights the importance of performance metrics in estimating football players' market values and provides a pathway for clubs to make more informed and data-driven decisions in the transfer process. These findings can make significant contributions to future research and practice in football economics and club management.

Limitations

While this study provides valuable insights into the role of performance metrics in estimating the market values of football players in Europe's top five leagues, certain limitations should be acknowledged.

- Dataset Scope: The dataset includes performance metrics and market values for 1508 players from a single season. This focus on a specific timeframe limits the generalizability of the findings to other seasons or leagues outside the top five in Europe.
- Market Value Subjectivity: Market values were obtained from the Transfermarkt platform, which partially relies on subjective inputs such as expert opinions and user votes. While these values are widely recognized, they may not fully capture market dynamics or fluctuations influenced by factors like recent transfers or injuries.
- Exclusion of External Variables: Non-performance-related factors such as player endorsements, media popularity, and off-field behavior, which can significantly impact market values, were not included in the model. Incorporating such variables could enhance the model's accuracy.

- Model-Specific Biases: The machine learning algorithms used (e.g., Random Forest and CNN) have inherent assumptions and biases, which might influence the results. Different algorithms or approaches could yield varying outcomes.
- Resource Constraints: Due to practical limitations, the study did not incorporate a longitudinal analysis to observe changes in market values over time, which might offer deeper insights into player valuation dynamics.

Despite these limitations, the study provides a robust framework for future research to refine and expand upon these findings. Incorporating larger, more diverse datasets and additional variables could further enhance the predictive accuracy and applicability of the proposed models.

CONCLUSION

This study has demonstrated the significant role of performance metrics in estimating the market values of football players in Europe's top five leagues. The findings suggest that football clubs can make more informed decisions during transfer processes when performance data is used accurately and effectively. Specifically, the high correlations between metrics such as the number of goals, shots per game, assists, man of the match, and age with player market values emphasize the importance of adopting a more analytical and data-driven approach in player evaluations.

Moreover, the high accuracy rates of the regression and classification models used in this study provide a valuable foundation for future research in football economics and club management. Prediction models based on performance metrics can help clubs optimize their transfer expenditures and ensure financial sustainability. In this context, our findings support football clubs in making strategic decisions that enhance the efficient use of financial resources, thereby aiding in their long-term success and sustainability.

PRACTICAL IMPLICATIONS

This study has significant practical implications for football clubs, analysts, and stakeholders in the transfer market. It demonstrates that integrating performance metrics into machine learning models provides a robust framework for estimating football players' market values. By adopting these models, clubs can make more informed, data-driven decisions during transfer negotiations, effectively reducing the risks of overpayment or undervaluation.

Identifying players whose performance metrics align closely with their market values allows clubs to optimize their transfer budgets. This approach is particularly beneficial for ensuring financial sustainability, especially for clubs with limited resources or those recovering from economic disruptions, such as the COVID-19 pandemic. Furthermore, metrics such as goals, assists, and passing accuracy offer an objective foundation for evaluating players' contributions, minimizing reliance on subjective assessments and biases. This improvement enhances the fairness and accuracy of talent evaluations.

The insights from this study also have the potential to revolutionize scouting processes. Clubs can prioritize performance metrics that show the strongest correlation with market value, enabling them to identify high-potential talent more effectively. Beyond football, the methodology proposed in this study could be adapted for other sports, such as basketball or baseball, or even extended to industries where performance metrics influence valuation, such as corporate talent assessments.

Although these implications are promising, their practical implementation requires access to comprehensive performance datasets and expertise in machine learning. Collaboration between football clubs and analytics providers will be essential to address these challenges, ensuring the proposed models achieve their maximum potential impact.

Acknowledgements

The authors declare that no financial support was received from any organization or institution for the research, authorship, and publication of this article.

Authors' Contributions

Both authors have equally contributed to all stages of this research. This includes the conception and design of the study, data collection, analysis and interpretation of the data, as well as the drafting and revision of the manuscript. Both authors read and approved the final manuscript.

Declaration of Conflict Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethics Statement

This study was conducted using publicly available online data pertaining to football players' performance metrics and market values. No experiments or tests involving human participants were carried out. Therefore, ethical approval from an ethics committee was not required for this research.

REFERENCES

Agresti, A. (2012). *Categorical Data Analysis*. John Wiley & Sons.

Al-Asadi, M. A., & Taşdemir, Ş. (2022). Predict The Value of Football Players Using Fifa Video Game Data And Machine Learning Techniques. *IEEE Access*, 10, 22631–22645. https://doi.org/10.1109/ACCESS.2022.3154767

Arrul, V. S., Subramanian, P., & Mafas, R. (2022, April). Predicting The Football Players' Market Value Using Neural Network Model: A Data-Driven Approach. *In 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICDCECE53908.2022.9792860

Aydemir, A. E., Taskaya Temizel, T., & Temizel, A. (2022). A Machine Learning Ensembling Approach to Predicting Transfer Values. *SN Computer Science*, 3(3), 201. https://doi.org/10.1007/s42979-022-01095-z

BBC Sport. (2012). Rangers Football Club Enters Administration. *BBC Sport*. Retrieved June 10, 2024, from https://www.bbc.com/news/uk-scotland-glasgow-west-17026172

Behravan, I., & Razavi, S. M. (2021). A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*, 25(3), 2499–2511. https://doi.org/10.1007/s00500-020-05319-3

Bida, M., & Mirzoyan, A. (2023). Factors Influencing Transfer Policy of Football Clubs. *Journal of the New Economic Association*, 58(1), 66–88. https://doi.org/10.31737/22212264_2023_1_66

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Carling, C., Williams, A. M., & Reilly, T. (2005). *Handbook of soccer match analysis: A systematic approach to improving performance*. Routledge.

Chai, T., & Draxler, R. R. (2014). Root Mean Square Error (Rmse) or Mean Absolute Error (Mae)? Arguments Against Avoiding Rmse in The Literature. *Geoscientific Model Development*, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Deloitte. (2021). *Football Money League 2021*. Retrieved May 12, 2024, from https://www2.deloitte.com/cn/en/pages/international-business-support/articles/deloitte-football-money-league-2021.html

Deloitte. (2024). *Football Money League 2024*. Retrieved May 20, 2024, from https://www.deloitte.com/uk/en/services/financial-advisory/analysis/deloitte-football-money-league.html

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). John Wiley & Sons.

Fawcett, T. (2006). An Introduction to Roc Analysis. *Pattern Recognition Letters*, 27(8), 861-874.

Franceschi, M., Brocard, J. F., Follert, F., & Gouguet, J. J. (2024). Determinants of football players' valuation: A systematic review. *Journal of Economic Surveys*, 38(3), 577-600. https://doi.org/10.1111/joes.12552

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in* R. Spinger.

Gazzetta dello Sport. (2015). *Parma Officially Bankrupt: All Employees Sacked*. Retrieved May, 20, 2024 from https://www.gazzetta.it/Calcio/Serie-A/Parma/11-02-2015/ecco-voragine-conti-parma-96-milioni-debiti-100846743143.shtml

Giulianotti, R. (1999). *Football: A Sociology of The Global Game*. Polity Press.

Goldblatt, D. (2006). *The Ball is Round: A Global History of Soccer*. Riverhead Books.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Herm, S., Callsen-Bracker, H. M., & Kreis, H. (2014). When The Crowd Evaluates Soccer Players' Market Values: Accuracy and Evaluation Attributes of an Online Community. *Sport Management Review*, 17(4), 484-492. https://doi.org/10.1016/j.smr.2013.12.006

Hudl, S. (2019). *Wyscout Forum: The Ultimate Football Network.* Retrieved June 1, 2024 from https://www.hudl.com/blog/wyscout-forum-the-worlds-1-event-in-football-for-transfer-negotiations

Inan, T., & Cavas, L. (2021). Estimation Of Market Values of Football Players Through Artificial Neural Network: A Model Study from The Turkish Super League. *Applied Artificial Intelligence*, 35(13), 1022-1042. https://doi.org/10.1080/08839514.2021.1966884

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4-37. https://doi.org/10.1109/34.824819

Kologlu, Y., Birinci, H., Kanalmaz, S. I., & Ozyilmaz, B. (2018). A Multiple Linear Regression Approach for Estimating The Market Value Of Football Players In Forward Position. *Arxiv Preprint* arXiv:1807.01104.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied To Document Recognition. *Proceedings of the IEEE, 86*(11), 2278-2324. https://doi.org/10.1109/5.726791

Lee, H., Tama, B. A., & Cha, M. (2022). Prediction of football player value using Bayesian ensemble approach. *arXiv Preprint*, arXiv:2206.13246.

Leifheit, N., & Follert, F. (2023). Financial Player Valuation from The Perspective of The Club: The Case of Football. *Managing Sport and Leisure*, 28(6), 618-637. https://doi.org/10.1080/23750472.2021.1944821

Li, C., Kampakis, S., & Treleaven, P. (2022). Machine Learning Modeling to Evaluate The Value Of Football Players. *arXiv Preprint* arXiv:2207.11361.

Liaw, A., & Wiener, M. (2002). Classification And Regression By randomForest. *R News*, 2(3), 18-22.

Mann, D. L., Dehghansai, N., & Baker, J. (2017). Searching For the Elusive Gift: Advances In Talent Identification In Sport. *Current Opinion in Psychology*, 16, 128-133. https://doi.org/10.1016/j.copsyc.2017.04.016

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction To Linear Regression Analysis* (5th ed.). John Wiley & Sons.

Morgans, R., Adams, D., Mullen, R., McLellan, C., McNaughton, L., & Williams, M. (2014). A Comparison of Physical And Technical Match Performance of A Team Competing In The English Championship League And Then Promoted To The English Premier League. *International Journal of Sports Science & Coaching*, 9(5), 1101-1114. https://doi.org/10.1260/1747-9541.10.2-3.543

Murray, B. (1994). *The World's Game: A History of Soccer*. University of Illinois Press.

Pariath, R., Shah, S., Surve, A., & Mittal, J. (2018, March). Player Performance Prediction İn Football Game. In 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1148-1153). IEEE. https://doi.org/10.1109/ICECA.2018.8474750

Powers, D. M. W. (2011). Evaluation: From Precision, Recall And F-Measure To Roc, İnformedness, Markedness And Correlation. Journal of Machine Learning Technologies, 2(1), 37-63. https://doi.org/10.48550/arXiv.2010.16061

Sokolova, M., & Lapalme, G. (2009). A Systematic Analysis of Performance Measures For Classification Tasks. *Information Processing & Management*, 45(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Transfermarkt. (2024). *Market Values*. Retrieved June 3, 2024, from https://www.transfermarkt.com/navigation/marktwerte

Whoscored. (2024). *Player Statistics*. Retrieved May 20, 2024, from https://www.whoscored.com/Statistics